
# Natural Language Processing – Assisted Extraction of Information from Literature

# Approach

- Utilize SRA's natural language processing tool (NetOwl® Extractor) on the ERIC-BRC literature space to speed extraction of meaningful information
  - SRA working on NLP since 1986, NetOwl first released by SRA in 1996.
  - Used by many fields including financial, intelligence.
  - Robust and well accepted (National Institute of Standards and Technology "Automated Content Extractions (ACE)" trials, top scorer for the past three years)
- Application for ERIC:
  - Create training and blind sets
  - Write patterns to extract information
  - Build interface so scientists can use information *to assist* annotation

# Information to Extract - entities

- Genes and gene products: (invC and InvC)
- Operons: (paaABCDE, inv locus)
- Organism: (*E. coli*, *Shigella flexneri*, human)
- Strain: (K-12, LT2)
- Non-enzyme protein: (prepilin, actin, gluconate transporter)
- Enzymes: (CDP-tyvelose epimerase, cysteine desulfurase)

# Information to Extract - links

- Relationships between entities and what the entities do.
- Three types
  1. (gene, gene product, operon) to role
  2. (gene, gene product, operon) mutation to result
  3. organism to pathogenesis
- Extraction based on syntactical structure

# Gene to Role

- "These results suggest that VirK function is an essential virulence determinant for shigellae involved in the expression of virG gene product at post-transcriptional level."
  - VirK - is an essential … post-transcriptional level
- "The sitA gene encodes a putative periplasmic binding protein, sitB encodes an ATP-binding protein, and sitC and sitD encode two putative permeases (integral membrane proteins)."
  - sitA - encodes a putative periplasmic binding protein
  - sitB - encodes an ATP-binding protein
  - sitC & sitD - encodes two putative permeases (integral membrane proteins
- "including catalase-peroxidase (KatY), murine toxin (Ymt)."
  - KatY - catalase-peroxidase
  - Ymt - murine toxin

# Mutation to Result

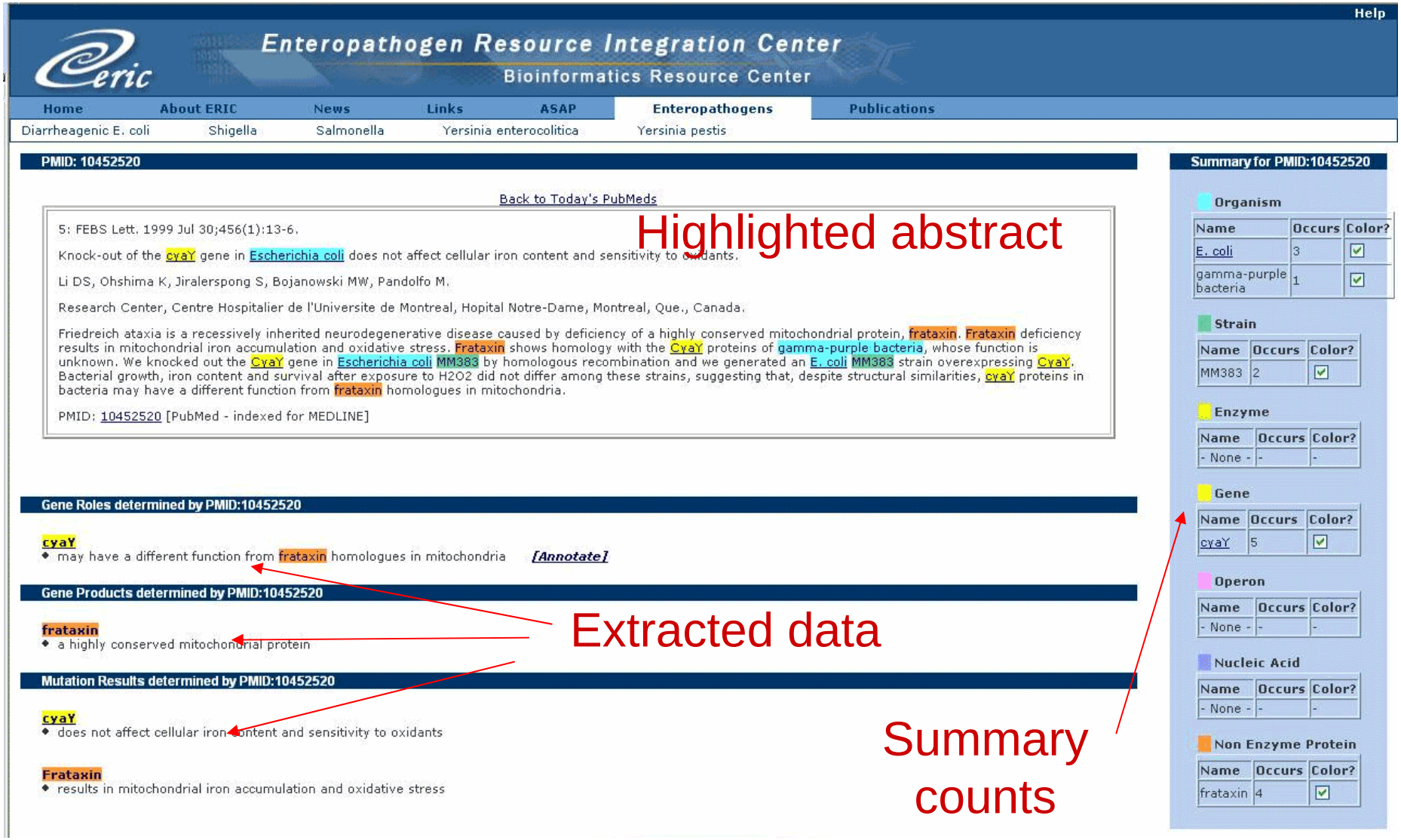
- "After inactivation of icsB, the mutant strain remained invasive, but formed abnormally small plaques on HeLa cell monolayers, colonized only the peripheral cells of Caco-2 islets, and was unable to provoke a keratoconjunctivitis in guinea-pigs."
  - icsB - the mutant strain remained invasive
  - icsB - formed abnormally small plaques on HeLa cell monolayers
  - icsB - colonized only the peripheral cells of Caco-2 islets
  - icsB - was unable to provoke a keratoconjunctivitis in guinea-pigs

- "The xdhA mutant grew faster with aspartate as a nitrogen source. The mutant also exhibited sensitivity to adenine, which guanosine partially reversed."
  - xdhA – grew faster with aspartate as a nitrogen source.
  - xdhA – exhibited sensitivity to adenine, which guanosine partially reduced

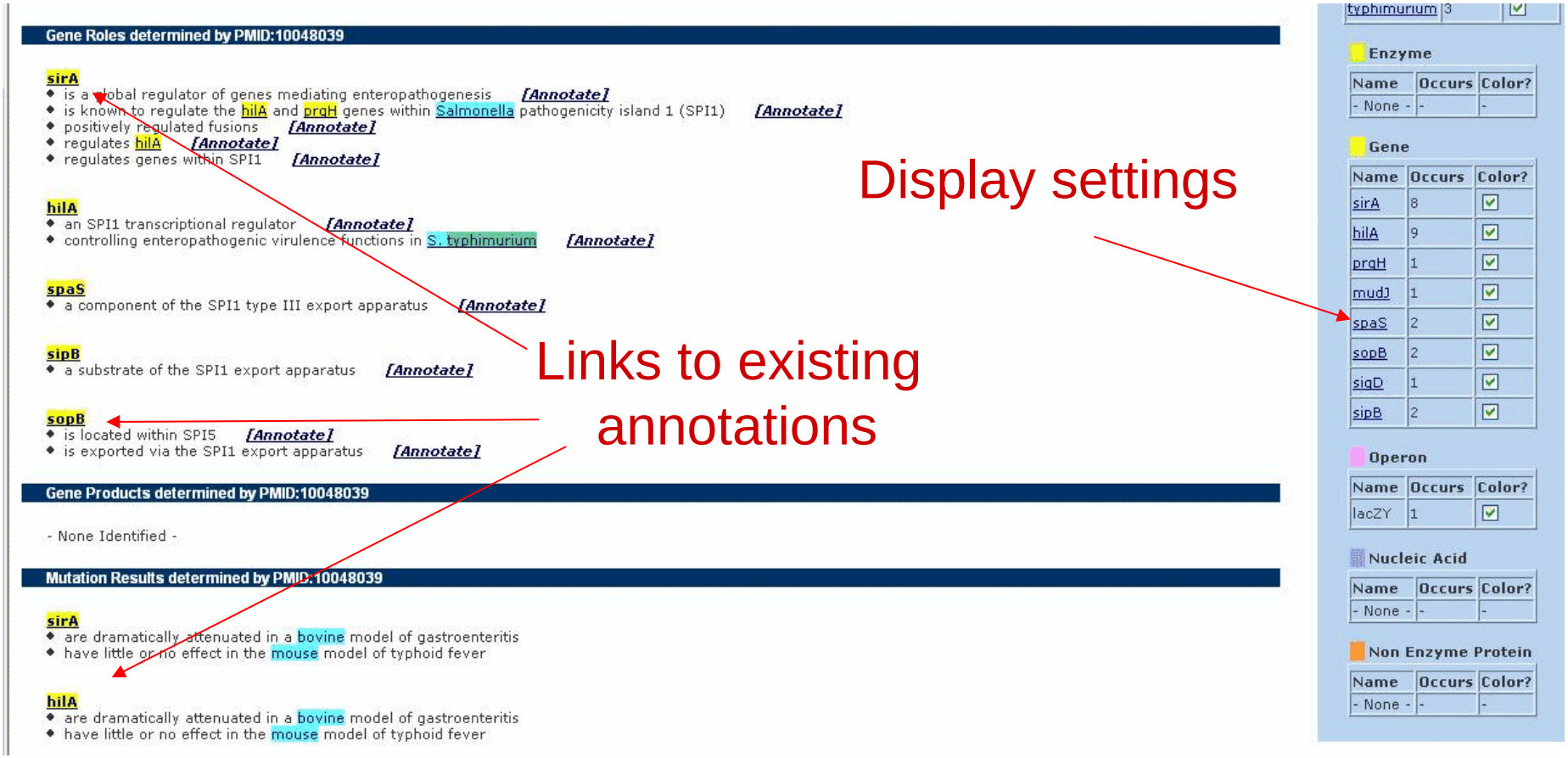
# Organism to pathogenesis

- "Because S. typhimurium causes gastroenteritis in both cattle and humans, we believe that this information may be directly applicable to the human."
  - S. typhimurium - causes gastroenteritis in both cattle and humans
- "Shigella flexneri causes bacillary dysentery by invading epithelial cells of the colonic mucosa"
  - Shigella flexneri - causes bacillary dysentery

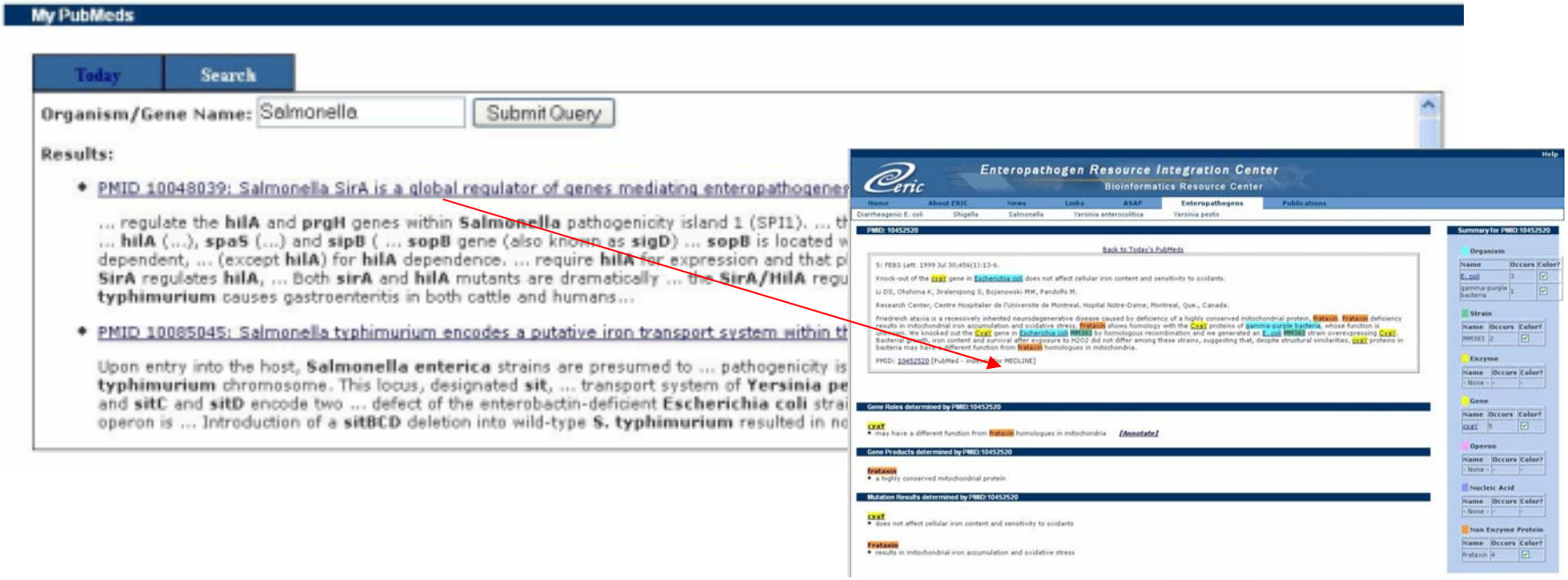
# A Proposed Use



# A Proposed Use - detail



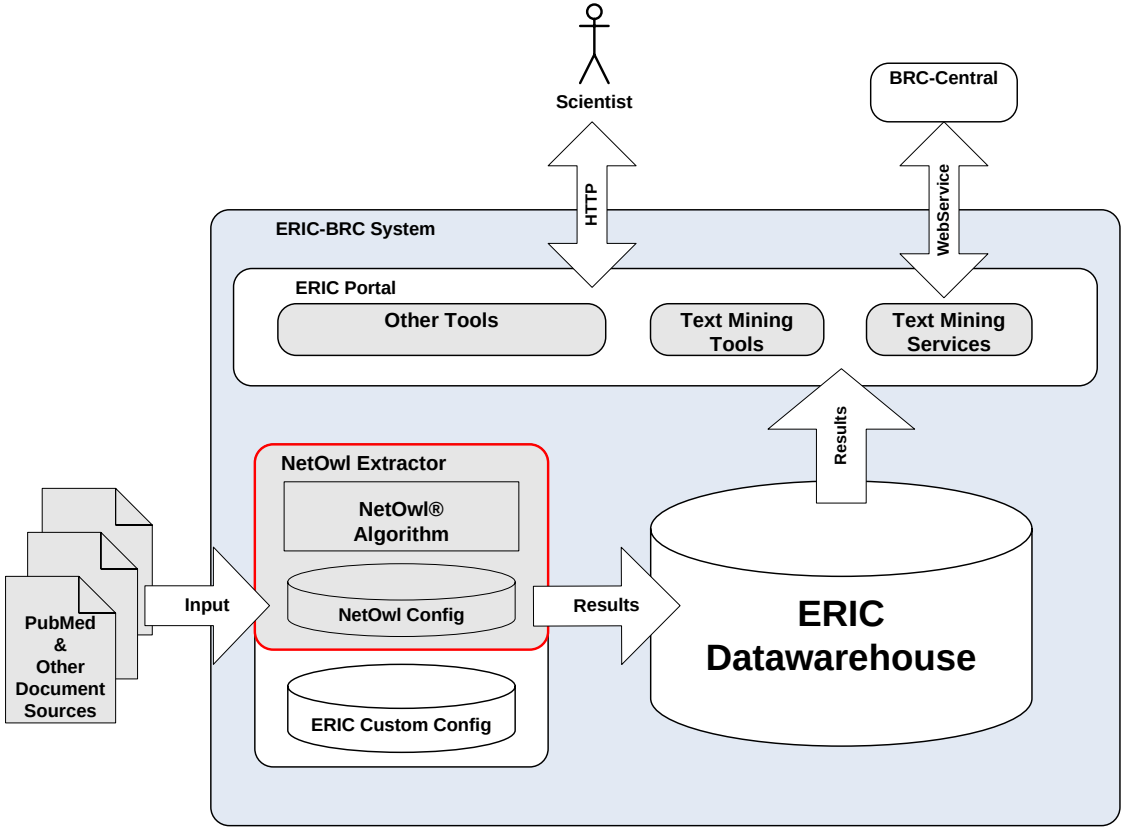
# Example Workflow – Examining Current Literature




- Today's relevant abstracts processed through system. Interface will present these results to scientific community allowing them to drill down to detail
- Interface links to existing annotations allowing user to determine what is new in this article, add and modify annotations
- Can be either all articles or from a set of keywords (eg. specific genes)

# NetOwl Extractor
## Optimizing Manual Literature Annotation

- NetOwl is privately developed software by SRA International.
- Being configured to mine information from abstracts and journals articles relevant to ERIC.
- Will increase efficiency by identifying interesting information for annotators.




*Software licensed for use on ERIC bounded in red.